Docket No. F-8015

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant	:	Wayne DAWSON, et al.
Serial No.	:	10/695,247
Filed	:	October 27, 2003
For	:	A METHOD FOR PREDICTING THE SPATIAL-ARRANGEMENT TOPOLOGY OF AN AMINO ACID SEQUENCE USING FREE ENERGY COMBINED WITH SECONDARY STRUCTURAL INFORMATION
Group Art Unit	:	1631
Examiner	:	Karlheinz R. Skowronek
Confirmation No.	:	5890
Customer No.	:	000028107

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

DECLARATION OF WAYNE DAWSON UNDER 37 C.F.R § 1.132

I, Wayne Dawson, declare and say:

I am a citizen of the United States and I reside in Japan.

I am one of the coinventors in U.S. Patent Application No. 10/695,247.

I graduated in 1989 from San Jose State University, in California, U.S.A, with a B.S. in physics. I also graduated in 1992 from San Jose State University,

in California, U.S.A., with an M.S. in physics. I also graduated in 1996 from the University of Tokyo, in Tokyo, Japan with a Ph.D. in physics.

I have been working in the field of the subject matter of U.S. Patent Application No. 10/695,247 for more than 10 years and have practiced computational physics in relation to structure analysis and prediction of both inorganic solids and organic molecules for about 20 years. In fact, I am one of the authors of Dawson et al., as cited by the Office Action of September 12, 2007 in U.S. Patent Application No. 10/695,247.

My current employment is at the Chiba Institute of Technology as a research associate. I have been employed by the Chiba Institute of Technology since 2004.

I am familiar with U.S. Patent Application No. 10/695,247 and the documents applied in the rejection of the claims of U.S. Patent Application No. 10/695,247 as explained in the Office Action of September 12, 2007.

I have analyzed both the invention of U.S. Patent Application No. 10/695,247 as well as the three documents (Floudas et al., Alm et al., and Dawson et al.) applied to reject the claims of U.S. Patent Application No. 10/695,247. I have some comments regarding the invention of U.S. Patent Application No. 10/695,247 and the three documents applied to reject the claims of that

application. Please find enclosed Appendix 1, Appendix 2, and Appendix 3 with my comments.

I further declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

Date February 11th 2008 By Wayne Dawson

Wayne Dawson

Enc. Appendix 1
Appendix 2
Appendix 3

APPENDIX 1

A) The entropy model and methods in Dawson et al [1] are incomplete and inadequate

What Dawson et al. could calculate in their JTB research paper [1] is the entropy *given* a structure. Dawson et al did not demonstrate that they could build a structure prediction program and they depended on a selected set of given structures to test their model.

Dawson et al. understood how the concept of a persistence length influences the *global* entropy. However, they did not understand that this must be accounted for at all levels of the calculation. They assumed, just like the authors of the sources where they obtained the structures to test their model, that they could manage to calculate the behavior of the RNA (and their claims about proteins) using a strictly monomer-by-monomer calculation scheme. They did not understand that they must fix their calculation strategy to account for the fact that the monomers (amino acids and nucleic acids) interact locally as a group in persistence-length related length-scales.

The persistence length introduces a straightening effect that requires constraints to account for it. These constraints reduce the number of degrees of freedom. In effect, the monomer-by-monomer unit approach must be replaced by a group of monomers: more of a group-by-group approach. If a polymer of N monomers has complete freedom of motion at the position of each monomer, then the polymer has N degrees of freedom. Roughly speaking this means that, for a given polymer with persistence length ξ and N monomers, the polymer has approximately N/ξ degrees of freedom. Dawson et al didn't understand that they must account for this effect by adding constraints on a local scale to their approach.

The only reason Dawson et al were successful in their study (which was directed to RNA, not proteins) was because the selective set of structures they obtained just so happened to avoid the pitfalls that their own approach would yield. Without applying the filtering and straightening effects of the persistence length, and the corresponding reduction in the degrees of freedom, they discovered that their structures were incorrectly predicted. In Figure 1, the correct prediction for a sequence of CUG repeats is shown using one of the sources of structures Dawson et al uses. Figure 2 shows what happened because Dawson at al only applied their global entropy prediction scheme without reducing the degrees of freedom and introducing constraints. The structure crinkles up because there are too many degrees of freedom and the flexibility of the structure is over determined. When Dawson et al. apply constraints and filter out the excess degrees of freedom; they again obtain the structure shown in Figure 1. However, this work was not reported until Feb 2006 and first submitted in July of 2005 [2]. The

RNA pseudoknot approach, from which they borrowed the main concepts of the model for protein structure prediction, was not reported until Sept of 2007 and submitted in July of 2007 [3].

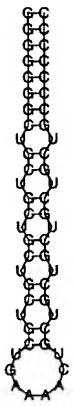

Figure 1. The calculated secondary structure of a CUG repeat sequence using a standard genre of RNA-structure-prediction programs. This structure prediction is essentially correct. It is also predicted using the current versions of vsfold4 and vsfold5 developed by Dawson et al. (refs [2] and [3]).

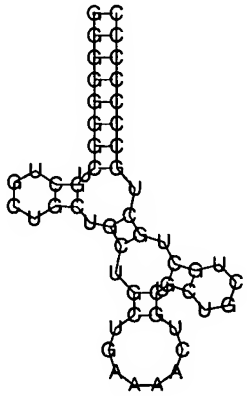

Figure 2. The calculated secondary structure of a CUG repeat sequence when the model of Dawson et al is used without correcting with constraints (straightening effects caused by the persistence length) and limiting the number of degrees of freedom. This structure prediction is incorrect and changing the persistence length to 10 nt or more does not change this result even though this should mean that the structure will tend to straighten over at least the distance of 10 nt.

Thus, the model reported in Dawson et al (ref [1]) does not solve the structure prediction problem. Dawson et al relied on other sources to supply the structures. These sources, due to their own idiosyncratic problems, just happened to have structure

sets that were far worse than the correct structures. This meant that Dawson et al were able to find more of the correct structures because their global approach was at least correct. However, when Dawson et al were given to using their technique to independently predict structures, they failed because they did not know that these erroneous approaches had fortuitously selected structures that agreed with the constraints Dawson et al needed.

This is not because the other methods knew anything about persistence length or reduction of degrees of freedom either. Rather, it was because of two errors that just so happened to cancel each other sufficiently that something intelligible was produced.

The first of these is that the other models violate the second law of thermodynamics (see Appendix 3). It therefore encourages a straightening effect as a consequence. It was not a reduction in the degrees of freedom, only ignorance about what entropy means. By neglecting the persistence length and violating the second law, they generate a set of structures that are often far worse than the correct ones; particularly when the sequences were long.

The second of these was that the particular sequences and correct corresponding RNA structures tended to be very stiff structures to begin with (having a long persistence length). The entropy model used by these sources is fairly good at finding long straight stem structures, and if these also happen to be the selected set of test structures, it is no surprise that at least some of the predicted results would agree with the experimental data.

Therefore this made it easy for Dawson et al to find the reasonably correct structures in the list. However, when Dawson et al's model was tested on its own, it also failed at the other extreme because they did not know that they had received a highly selective set of structures until that point. Only when Dawson et al added these constraints on the degrees of freedom were they able to conquer the RNA structure prediction problem in a significant way [2-3].

The issue discussed here applies to all polymer problems including proteins. What Dawson et al learned from the example in Figure 2 is that constraints have to be introduced to limit the amount of variation and the degrees of freedom.

Therefore, the discovery that the structure predictions require straightening and structural constraints and, in particular, filtering of the degrees of freedom to properly express the straightening effects of polymer behavior, was not anticipated by Dawson et al. They did not test their idea and, with what they provide, they cannot account for this problem in their procedures without being given a selected set of other structures from a different approach that does not offer these crinkled up structures.

Moreover, their model only handles RNA secondary structure. Secondary structure requires that for any base pair indices i and j with $i<j$, all other base pairing (i' and j' with $i'<j'$) must satisfy either $i<i'<j'<j$ or $i',j'<i$ or $i',j'>j$. RNA pseudoknots allow that this rule can be broken such that $i'<i<j'<j$ or $i<i'<j<j'$. Folded beta sheets and alpha helices have repetitive looping and pairing indices that resemble RNA pseudoknots. The analysis procedures of Dawson et al cannot correct RNA pseudoknot predictions from other sources without introducing new techniques that they did not publish until September of 2007 [3].

Finally, the fact that entropy applies to all matters of engineering and physics does not in of itself anticipate anything. Whereas Dawson et al can realize that their model could be applied to protein problems, Dawson et al did not provide a method to do this calculation and in the case of arranging beta-strands into beta-sheets and arranging alpha-helices into an organized structure, Dawson et al did not publish any kind of technique that could achieve this type of art until September of 2007 [3].

The model proposed in U.S. Patent Application No. 10/695,247, which was later used to solve the RNA pseudoknot problem, is unique. No other method uses such mapping, group reduction and filtering of degrees of freedom to reflect the persistence length and sound thermodynamic rules to arrive at the predicted structures. They all rely on fallacious thermodynamic models that generally violate the second law of thermodynamics (Appendix 3).

B) Objections to cited Reference 4 (Floudas et al).

A persistence length of 5 amino acids (AA) seems to be rather common in proteins and RNA. However, the reason for the use of the AA fragments of length 5, 7, 9 etc. is not designed to account for structures that have longer or shorter persistence lengths. Indeed, neither the persistence length nor the Kuhn length is even mentioned in Floudas et al. Moreover, the claims in Floudas et al. are only applied to fragments of alpha-helices, not to beta-strands.

For beta-strands, Floundas et al. mention "constraints" and "restraints" (discussed intermittently from column 25, line 14 to column 31 line 36). The proposition involves forcing specific Ramachandran angles (i.e., a beta-strand configuration) on either an individual amino acid or a group of amino acids in a molecular dynamic simulation. It is a property that is neither distributed nor assigned throughout the amino acid sequence and is not used to define loops or turns. As with the alpha-helices, the reason for

applying these restraints to the AA fragments is not determined by the persistence length or the Kuhn length. The application is applied to specified parts of the sequence as a tool.

The current embodiment handles the secondary structure of the protein as a group and assumes that the AAs fold as a group. The model depends on the persistence length as this effect changes the degrees of freedom. The results do not depend on the length of the beta-strand region, only the persistence length and the results of the calculated free energy. There is no such concept in Floundas et al (reference [4]).

Molecular dynamics (MD) simulations techniques currently lack the ability to model this persistence length change that can occur between AA sequences in a given set of solvent conditions. They do account for the self interaction between neighboring amino acids. However, this only yields a persistence length of the order of 2 or 3 [AAs] whereas the general behavior of these materials is much closer to 5 [AAs] and sometimes far more. Double strand DNA and RNA can have persistence lengths of 150 nucleotides, and one would expect that a continuous alpha helix of length 1000 could have a persistence length of about 40 [AAs] or more. MD simulation appears to remain poor at anything modeling more than the self interactions between the atoms in a system. Perhaps this is due to the interactions between water and the polymer. The water would be expected to form a crystalline structure around the polymer, particularly if the monomer is not highly water soluble. There may be effects from quantum mechanics that dominate in water-water, water-ion, and water-polymer interactions that are longer range than is currently well modeled with MD simulation. At any rate, it appears to be a non-trivial problem in polymer-solvent-ion modeling and this is what makes first-principles-modeling of beta-strands interactions using MD simulation extremely difficult.

The methods in reference [4] do not use a method of crawling along the sequence of secondary structure testing for the folding likelihood and weighting the order of folding with the distance in sequence length between the different points of secondary structure. They are also more directed to finding the specific local arrangement rather than how these larger blocks come together globally; though obviously, this was part of their described goal.

While it may be the Examiner's position that it has been attempted to look for the topology in the prior art, Applicants respectfully dispute that the prior art tried to search for that topology as in the present invention.

C) Objections to cited Reference 5 (Alm et al.)

First, the model that reference 5 uses only calculates the entropy change for closing a loop (page 11306, column 1, middle of third full paragraph: or line 54 from the top). Their L_o represents the closing a loop of length L_o, not the total entropy loss due to forming structured regions of beta-strands into beta sheets. It is not the full contribution of free energy.

Second, reference [5] utilized a two state system: one state was the denatured configuration (random) and the other was the native state. This is effectively a generalized lattice model where each configuration has two possible states (which qualify as generalized coordination number q). Indeed, the authors specify that the model has 2^N configurations (page 11305, column 2, end of paragraph 2). This implies a model of the form q^N configurations, where q is a constant.

Lattice models have a greatly disguised constraint (the lattice). This adds a highly restricted spatial orientation that is only approximately true.

Lattice models also have a pitfall within the model itself (see the first part of Appendix 2). When the number of states (the coordination number of the lattice) is larger than the number of monomers ($N < q$), a lattice model will over-predict the number of configurations. For a two state system this is not so serious because it is the smallest coordination number. However, when the number of monomers greatly exceeds the specified number of states, the number of configurations is greatly *underestimated*. The number of configurations does not satisfy a Gaussian (or freely jointed) polymer chain model (Appendix 2). No one here disputes that the closing entropy for a loop should be in some way roughly proportional to $\ln(L_o) + \text{constant}$, as presented in reference 5 equation 1. Yet Equation 1 (reference 5) is based on a Gaussian model. Both answers cannot be true unless they are consistent, but we show in Appendix 2 that the lattice model leads to a contradiction.

For example, since no constraints of any kind are specified in these models, one should be able to predict the same number of configurations from a lattice model as they do from a Gaussian model. The correct conformation pattern should follow a rule that corresponds to the following expression

$$C_N \approx \left(\frac{q(N)}{w}\right)^{\alpha(N)} g(N)^{\beta} \qquad (1)$$

where $q(N)$, $g(N)$ and $\alpha(N)$ are increasing functions of N, and w and β are a

constants. If $q(N)=\Psi N$, $g(N)=\Psi N$, $\alpha(N)=\gamma N$, $w=e^{2\gamma}$ and $\beta=\gamma/\Psi$, then we obtain the following,

$$C_N \approx \left(\frac{\Psi N}{e^{\gamma}}\right)^{\gamma N} (\Psi N)^{\gamma/\Psi} \tag{2}$$

The logarithm of this is

$$\ln(C_N) \approx \gamma N \ln(\Psi N) - 2\gamma N + (\gamma/\Psi)\ln(\Psi N) \tag{3}$$

and the derivative with respect to N is

$$\frac{\partial(\ln(C_N))}{\partial N} \approx \gamma\left\{\ln(\Psi N) - 1 + \frac{1}{\Psi N}\right\} \rightarrow \Delta S(N) \approx \gamma k_B \left\{\ln(\Psi N) - 1 + \frac{1}{\Psi N}\right\} \tag{4}$$

which yields the equation used in Dawson et al.

For reference, when this equation uses the values $q(N)=N$, $g(N)=N$, $\alpha(N)=N$, $w=e$ and $\beta=1/2$, we obtain

$$C_N \approx \left(\frac{N}{e}\right)^{N} (N)^{1/2} \tag{5}$$

which is very close to the asymptotic approximation known as Stirling's formula

$$N! \approx (2\pi)^{1/2}(N/e)^N N^{1/2}. \tag{6}$$

where $N!=1\cdot 2\cdot 3\cdots N$. The total number of ways one can arrange N distinguishable objects is also $N!$ in size. All the amino acids in a protein (or RNA) are (in principle) distinguishable in the art of X-ray crystallography or NMR spectroscopy. They also tend to obey Maxwell-Boltzmann statistics which are used when computing the statistics of distinguishable objects (known as "particles" in the art of physics). Therefore their conformations must also be of order $N!$ in size.

It is shown in Appendix 2 (Section A2.2) that the summation of base pairing is a form of integration and that the actual integral of this argument leads to results that resemble Equation (3) where it is noted that the number of conformations is of order

N^N. From Equation (6) and (7), we can see that this is of similar size to a factorial expression. Hence, the CLE model is at least self consistent at this level.

Also, the true conformation limits on folding a beta-strand back and forth can be largely accounted for by including a weight δ on the logarithm term of Equation (4). A correct generalization of these equations is written in reference [2].

On the other hand, the lattice model assumes that $q(N) = \text{constant} \equiv q_o$ and, as expressed in reference [5] (page 11305, column 2), often assume that $g(N) = 1$ and $\beta = 1$. With the exception of a range of values around q_o, this does not match conceptually with a simple integration of Equation (4) where integration clearly projects some reasonable order of magnitude to the number of conformations available to such a structure. Neither does it return anything remotely resembling a Gaussian model when we try to evaluate its derivative: $d\left[\ln(2^N)\right]/dN = \ln(2)$.

Since typical computations of these configurations have been restricted to $4 < N < 20$ due to the extreme computational demands, it should be little wonder that this problem has been disguised. Nevertheless, it has been a mystery why we have these two incompatible conclusions. Whatever model we may chose to use, we must require that we obtain an answer that is consistent with other models purporting to do the same job. Equations (1)-(4) given above are consistent with a Gaussian-type model.

The lattice model with $q(N) = q_o$ should be weighted by the degeneracy $\sigma(N)$ of the configurations

$$\sigma(N) = \frac{q(N)g(N)^{\beta/\alpha(N)}}{q_o w} \tag{7}$$

to avoid neglecting the fact that we generate too many states when $N < q_o$ and too few when $N > q_o$ (Appendix 2). This degeneracy correction in Equation (7) reveals an increasing function of N ($q(N)$), a root mean square deviation (a measure of the variance), and is weighted by a scaling factor w. In Equation (5), we can see that the coordination number is $q(N) = N$, the root mean square deviation expands to $g(N)^\beta = \sqrt{N}$ and the scaling factor is the exponential base $w = e$. A standard Gaussian distribution shows this variance. Again, the CLE model (also a Gaussian distribution) remains self-consistent with the number of conformation, its derivatives and its integrals.

Finally, Alm et al. start with structures that already have a given (known) native state. If the structure is already known, and we simply introduce a lattice of correct

structure choice per amino acid plus one erroneous option (something else), we are importing a lot of additional information that has little to do with folding or structure prediction.

Therefore, the model in reference [5] is not a proper entropy model, it does not take into account constraints due to grouping of amino acids and the persistence length, and it is a lattice model that disguises these constraints and improperly accounts for the configurations of the structure. Ref. [5] does mention reducing the number of configuration by a rule that involves grouping (page 11305, column 2, third full paragraph, second sentence, under "Enumerating configurations"), but they base it on no physical justification such as the persistence length. Furthermore, the rule is based on a two state model of helix-coil transitions. Such a rule does not account for beta strand formations; only helix-coil transition. Their work was largely directed at alpha helices as was the case with reference [4].

D. Mistaken definitions in Floudas et al., Alm et al., and Dawson et al.

In the reference to polynomial time (claims 2 and 7), the small n pertains to secondary structure, not N, which is the total number of amino acids in the protein sequence. This n is not something disclosed by these other methods in references [4] or [5], nor is it claimed in Dawson et al [1].

References:

1. Dawson, W., Suzuki, K. and Yamamoto, K. (2001) A physical origin for functional domain structure in nucleic acids as evidenced by cross-linking entropy. *J Theor Biol.* 213, 359-386 and 387-412. (article)
2. Dawson, W., Fujiwara, K., Futamura, Y., Yamamoto, K., and Kawai, G. (2006) A method for finding optimal RNA secondary structures using a new entropy model (vsfold). *Nucleosides, Nucleotides, and Nucleic Acids* 25, 171-189. (article)
3. Dawson, W., Fujiwara, K., and Kawai, G. (2007). Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. *PLoS One*, 2, 905. http://www.plosone.org/doi/pone.0000905.
4. Flondas et al. (2001) US Patent #6,832,162 B2 (Dec. 14 2004: submitted Feb 16 2001).

5. Alm E and Baker D (1999) Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. Proc. Natl. Acad. Sci. USA, 96: 11305-11310.

Appendix 2: Gaussian polymer chain statistics

The Gaussian polymer chain is based on the statistics of a freely jointed polymer chain in which each link is free to rotate over the entire 4π solid angle (able to rotate through every angle of latitude and longitude including back on itself, which even a real chain cannot do). Studies of the properties of a freely jointed polymer chain lead to the recognition of Gaussian statistics.[1, 2, 3] Thus, the freely jointed polymer chain is the *model* and Gaussian statistics are one of the *properties* of a particular model of the freely jointed polymer chain.

In Section A2.1, we show that the configurations of a freely jointed polymer chain model (the Gaussian polymer chain) cannot be universally modeled *for all given sequence lengths* using a single coordination number as used in lattice models. In Section A2.2, we show one of the ways that the summation rule used in the cross linking entropy model can be derived. There are actually more than four independent ways to derive this model, two of which were presented in Dawson *et al.*[4a] Here we present what is probably one of the most intuitively clear derivations. The Section ends with some considerations in the broader context.

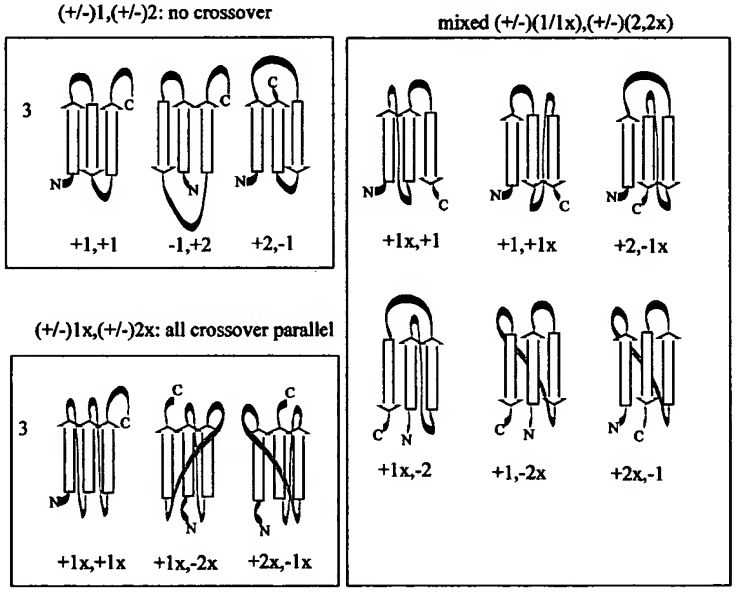

Figure A2.1. The full number of arrangements that can be generated from a protein composed of 3 beta-strands when crossover of the beta-strands is included. The notation below indicates the location of the next strand ($\pm1, \pm2$) and the x (±1x , ±2x) indicates a crossover beta-strand.[5] The total number of arrangements follows the rule $2^{n-1}n!/2$,[6] where n is the number of beta-strands.

A2.1. The freely jointed polymer chain model and the lattice model

Because it is difficult to do computer simulations on a true *freely-jointed* polymer chain, most work has been done using lattice models. However, is the lattice model a one-to-one representation of the freely jointed polymer chain? We find that the *freely jointed* polymer chain model (FJC-model) is expressed by a lattice model only if the scale of the coordination number (q) in the lattice model is similar to the number of residues in the polymer chain.

To show this, we turn to the example of protein beta sheets (β-sheets) because the concepts of protein structure are likely to be more familiar to most readers. However, we first must show that an equivalent β-sheets-like configuration can be generated with RNA pseudoknots.

The full range of configurations for a set of 3 beta-strands (β-strands) is shown in Fig A2.1. A set n β-strands (without other protein secondary structure) yields a total of $2^{n-1}n!/2$ unique β-sheets.[6] The β-sheets are grouped according to whether they possess crossover parallel β-sheets or no crossover. The notation follows that first used by Richardson *et al.*[7] The notation below indicates the location of the next strand $(\pm 1, \pm 2)$ and the x $(\pm 1x, \pm 2x)$ indicates a crossover β-sheets.

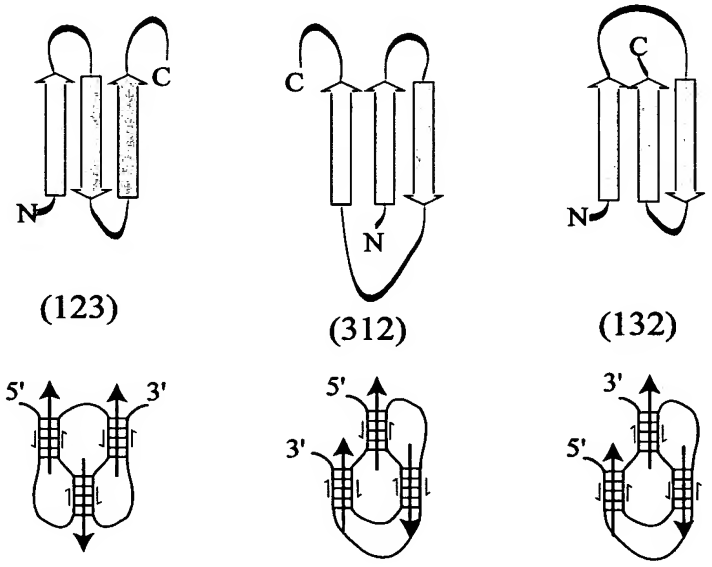

Figure A2.2. A one-to-one comparison between the patterns generated by 3 protein β-strands and an equivalent arrangement of RNA stems in the form of various RNA pseudoknots. The RNA cannot form a direct neighbor as happens with the beta-strands; however, by shifting them in the pattern of a ABACBC type pseudoknot (and other patterns), a similar pattern of structure can be found. The red stems are suggestive of a possible linkage stem (see Supplement).

In Fig A2.2, A pattern of 3 stems forming various RNA pseudoknots is compared with the equivalent pattern of β-sheet patterns *that contain no crossovers*; Fig A2.1 box with heading "no crossover". The arrows point along the 3' direction of the stem. Comparing the two patterns, they are effectively equivalent. The color coding on the stems is meant to imply the linkage stem (or stems) and the root domain (or domains). The color coding was discussed in Supplement 1 and 2. To some extent, there are some equivalents of crossovers that might be classified as orientation strain in Supplement 2; however, unlike proteins, they do not represent a unique topology. Hence, we focus on this smaller subset displayed in Fig A2.2. The combinatorics of this subset of protein beta-sheet structures (containing no crossovers) follows a $n!/2$ rule.

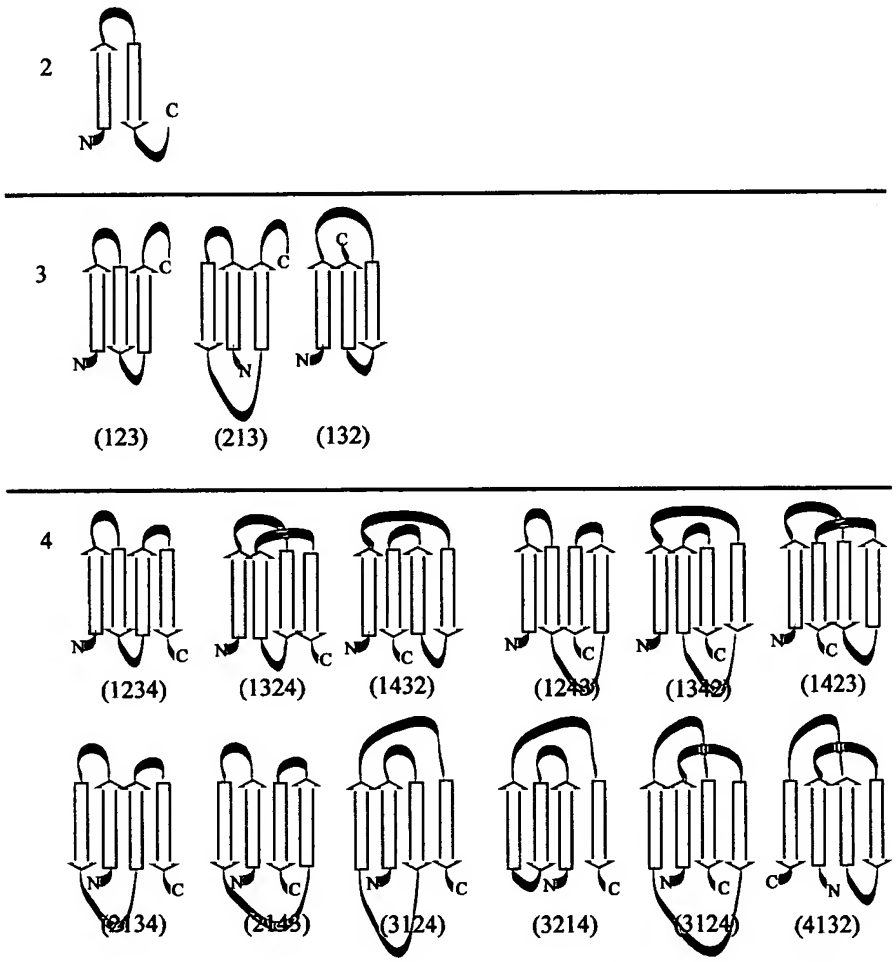

Figure A2.3. The number of unique arrangements of beta-sheets (excluding crossover parallel beta-sheets) for $n = 2,3,4$, where n is the number of beta-sheets. Because of a plane of 2-fold symmetry, the arrangements show factorial increase of $n!/2$.

Therefore, with no loss of generality in using protein beta strands, we consider a protein structure Ramachandran plot. We can divide the Ramachandran plot into three major regions α_R (alpha-helix), α_L (left-handed alpha-helix) and β (beta-sheet) sectors. Then the coordination number (q) is 3 and we would say that the number of possible configuration for an N residue protein is 3^{N-1}. However, this is not the only way we can cordon off the structure angles. For example, it would be reasonable to divide the beta-sheet regions into parallel- ($\beta_{\uparrow\uparrow}$) and anti-parallel ($\beta_{\downarrow\uparrow}$) beta-sheets, triple-helix (β_T) and polyproline [8] conformations ($\beta_{\mathrm{Pr}II}$). The coordination number is now increased to 6 and the N amino acids (aa) has 6^{N-1} conformations available to it. We could also add 3_{10} helices and beta-turn parameters to the list. Indeed, we could choose many additional ways to cordon off the Ramachandran plot in some mutually exclusive set of sectors. The coordination number is not unique because a lattice model constrains the FJC to a fixed coordination number.

Furthermore, it is easy to see that, for $N < q$, the configurations can overestimate the maximum number of configurations of N free particles ($N!$). For example, let $N = 7$ and $q = 8$, where we chose α_R, α_L, $\beta_{\downarrow\uparrow}$, $\beta_{\uparrow\uparrow}$, β_T, $\beta_{\mathrm{Pr}II}$, 3_{10} and beta-turn – all reasonable choices for a 7 monomer protein. Then $6! = 720$ but $8^6 = 262144$. The fixed coordination number far exceeds the configurations for N free particles.

It is much less apparent to realize that for $N >> q$, the fixed lattice model should *underestimate* the number of conformations accessible to free particles. To show this, we return to Fig A2.2 and consider the ways we might arrange an equivalent subset of β-sheets in a protein (Figure A2.3). The total number of ways that n β-sheets can be arranged is [6]

$$\Omega_\beta = 2^{n-1}(n!/2) \qquad \text{(A2.1)}$$

where the factor 2 accounts for the β-sheets that contain cross-over β-strands. The remainder ($n!/2$) is the combinatorial patterns of parallel and anti-parallel β-sheets with no crossover. Suppose we only permit combinatorial beta-sheet patterns that contain no crossover (Fig A2.1) and we make sequences in which the average β-sheets and turn segment is 6 residues. This means that we can write $n = N/6$ and we can compare it with the total number of configurations. Since the arrangement of the β-sheets depends on the number of conformations, the total number of β-sheets arrangements must not exceed the total number of conformation and, indeed, should be much less than that

$$n!/2 = (N/6)!/2 \ll 3^{N-1}, \qquad \text{(A2.2)}$$

where we assume the nominal coordination number $q = 3$. Taking the logarithms of Eqn (A2.2) and applying Sterling's approximation

$$\ln(n!/2) \approx \ln(\sqrt{2\pi}) + \left(n + \tfrac{1}{2}\right)\ln n - n - \ln(2). \tag{A2.3}$$

Rearranging and taking the limit on the dominant terms (with $n = N/6$) yields the following inequality

$$\lim_{N\to\infty} \frac{\left(\frac{N}{6} + \frac{1}{2}\right)\ln\left(\frac{N}{6}\right) - \frac{N}{6}}{N-1} = \frac{1}{6}\left(\ln\left(\frac{N}{6}\right) - 1\right) \ll \ln 3 \quad ?!? \tag{A2.4}$$

Eqn (A2.4) cannot be true for all N given a finite segment length containing a beta-strand and turn length (here a sum length of 6 aa). Solving Eqn (A2.4) yields $\ln N = 6\ln 3 + \ln 6 + 1$, or $N = 11980$ aa. This is a very large number, but increasing the segment length to any finite value will still not satisfy Eqn (A2.4) for *all* N. Nor will any finite q eliminate this discrepancy. Nor will including the prefactor 2^{n-1} in Eqn (A2.1) satisfy Eqn (A2.4), though N will become smaller. Therefore, we have a contradiction. Moreover, with factorial growth in the arrangement of beta-strands, the total number of conformations must also be of order $N!$ or $q \sim O(N)$.

At best, the coordination number is not a good measure of the true configuration space. Lattice models were intended for crystals where packing certainly defines and limits the orientations. However, in a FJC, the lattice only serves as a construct. The number of conformations of a freely jointed polymer chain cannot be universally estimated for all N using a single fixed coordination number (though it may approximate the number *for some* N). Further, q is not unique and lattice models *should include the degeneracy* in the evaluation.

A2.2. Cross linking entropy and the freely jointed polymer chain

Here we show that the coordination number for a polymer chain of N segments can be expressed consistently in a FJC when N and q are of the similar order; *i.e.*, the coordination number (q) is a function of N ($q = f(N)$). In other words, it is of similar order to the maximum number of conformations (order $N!$).

First we consider the defined parameters in a Gaussian polymer chain (see ref [4a&b]). The extensive parameters are the root-mean-square (rms) end-to-end distance (r) and the force (f) acting on the terminal ends of the polymer chain. From the definitions, the heat flow due to the work done by a polymer chain consisting of N monomers with state parameters r and f (in a

reversible reaction) is

$$TdS = dU + fdr\,, \tag{A2.5}$$

where U is the internal energy and S is the entropy. For a polymer, $\Delta U \sim 0$.[3] Eqn (A2.5) takes a form similar to the work done by an ideal gas in which fdr replaces pdV (where V is the volume and p is the pressure). The Helmholtz free energy is

$$F = U - TS\,, \tag{A2.6}$$

and making use of (A2.5), we get $dF = fdr - SdT$. Therefore,

$$S = -\left(\frac{\partial F}{\partial T}\right)_r \text{ and } f = \left(\frac{\partial F}{\partial r}\right)_T, \tag{A2.7}$$

whereupon $\partial S / \partial r = -\partial f / \partial T$. Since, from the definition of the GPC (see refs 4a&b), $S(r)$ is independent of T, $f = -T(\partial S / \partial r)$.

For the state parameters r and f

$$dS = \frac{\partial f}{\partial T}dT + \frac{\partial S}{\partial r}dr \tag{A2.8}$$

and if T is constant:

$$-TdS_T = -T\left(\frac{\partial S}{\partial r}\right)dr_T = fdr_T\,. \tag{A2.9}$$

From reference 4b with $\delta = 2$, $\gamma \geq 1$, and $\nu = 1/2$, the force equation comes from evaluating the derivative of the entropy:

$$f(r) = -T\left(\frac{\partial S}{\partial r}\right)_T = 2k_BT\left(\frac{\gamma}{r} - \alpha r\right). \tag{A2.10}$$

where the minimum is located at $r_o = \sqrt{\gamma/\alpha}$. This expresses the minimum in the end-to-end separation distance (not the rms-distance $\langle r^2 \rangle = \xi Nb^2$; ref. 4b). When $r < r_o$ or $r > r_o$, a force

drives the end-to-end distance back to r_o. For a GPC, $r_o = \sqrt{3\langle r^2 \rangle / 2}$.

So far, we have only considered the end-to-end distance. However, Eqn (A2.10) is not restricted only to the ends of the chain. For any (reasonable) number of residues (N) in a polymer chain, this same $\langle r^2 \rangle = \xi N b^2$ relationship holds. First, from reference 4b, the relationship can be translated. Second, for every length of sequence, this property holds. Hence, it is a general rule that applies to every point on a polymer chain. In short, for any indices i and j, where $i < j$, $\langle r^2 \rangle_{ij} = \xi (j - i + 1) b^2$.

We now ask what happens if we are to chose a set of specific coordinate pairs on the chain and apply a constraining interaction on them. The force equation resembles the displacement of a spring. Figure A2.4 shows a bank of 5 springs aligned in parallel to which a force is applied. The effective spring constant for an array of springs in parallel is the sum of the individual spring constants. Hence, the effective force is an additive property of the effective spring constant times the displacement

$$f = f_1 + \ldots + f_n = (k_1 + k_2 + \ldots + k_n)\Delta x \qquad \text{(A2.11)}$$

where Δx is displacement $k_1, k_2, \ldots, k_n$ are the individual spring constants, and $f_1, f_2, \ldots, f_n$ the individual contribution of spring k_l $(l = 1,2,\ldots n)$ to the force.

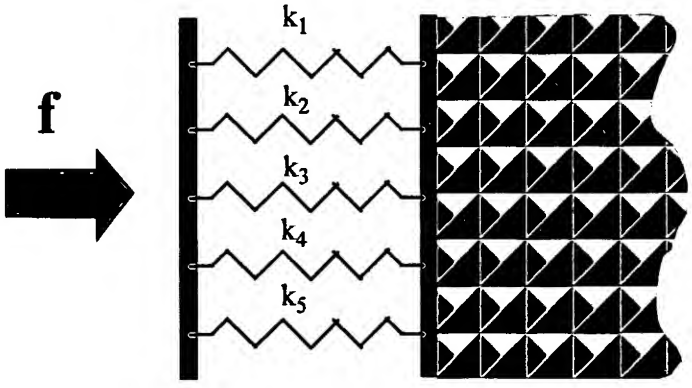

Figure A2.4. Example of a group of springs arranged in parallel with a force applied along the axis of the spring: k_l $(l = 1,\ldots 5)$. On the right hand side as a wall and a force f is applied from the left hand side. The response of these parallel springs is the sum of their spring constant.

By analogy, we build a similar-looking model in which a polymer chain is folded out to its equilibrium configuration and held in place by an array of springs (Figure A2.5). Just as pressure pushes against the surface of a container (force/area), so the equilibrium condition for the motion of the residues push and pull the contour of the polymer chain back to the equilibrium state. If we now force interaction between any pair of residues k, we observe a force f_k in response. The dependence of i and j on $r_k (= r_{ij})$ is only with respect to the number of residues separating them, and there is no explicit dependence of the other residues on the value of r_{ij} (a general characteristic feature of models like the GPC). Given this behavior, it follows that when n such forces are applied, we should expect a similar expression to emerge: namely,

$$f = f_1(r_1) + f_2(r_2) + \ldots + f_n(r_n). \qquad \text{(A2.12)}$$

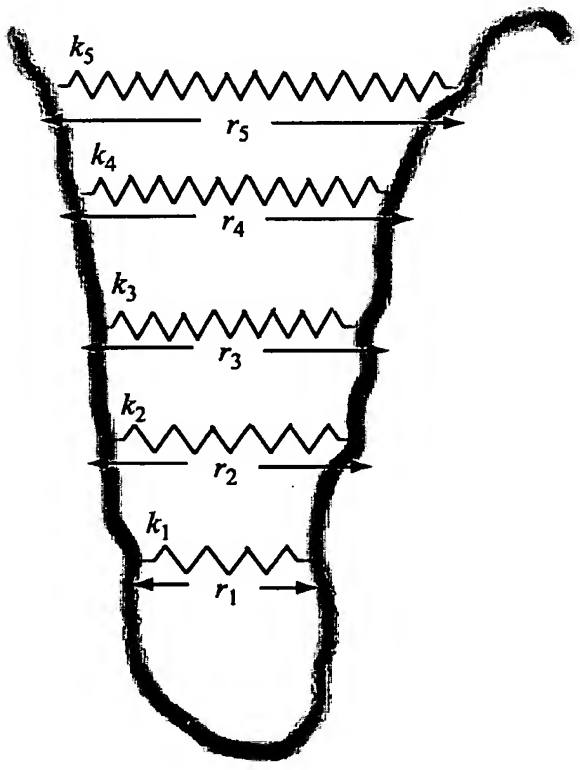

Figure A2.5. Based on the analogy presented in Figure A2.4, a group of springs arranged in parallel and set in equilibrium on a polymer chain.

Likewise, using Eqn (A2.9) for constant temperature, a displacement will be

$$TdS = T\left\{\left(\frac{\partial S}{\partial r_1}\right)dr_1 + \left(\frac{\partial S}{\partial r_2}\right)dr_2 + \ldots + \left(\frac{\partial S}{\partial r_n}\right)dr_n\right\} \tag{A2.13}$$

and

$$\int f dr = -\int TdS = \int f(r_1)dr_1 + \int f(r_2)dr_2 + \ldots + \int f(r_n)dr_n \tag{A2.14}$$

or, factoring out constant T

$$\Delta S = \Delta S_1 + \Delta S_2 + \ldots + \Delta S_n = \sum_l \Delta S_l \,. \tag{A2.15}$$

From the derivation of each ΔS_l (reference 4b), ΔS_l represents the probability that state l in configuration $r_{i,l}$ should acquire a configuration $r_{f,l}$: $p(r_{f,l} \cap r_{i,l})\Delta r$. The entropy $S_l(r_l)$ (reference 4b) corresponds to the probability of the configuration $r_{i,l}$ for state l: $p(r_{i,l})\Delta r$. The ratio forms a conditional probability

$$p(r_{f,l} \mid r_{i,l}) = \frac{p(r_{f,l} \cap r_{i,l})}{p(r_{i,l})} = \frac{p(r_{f,l})}{p(r_{i,l})}. \tag{A2.16}$$

Then, writing this in terms of the entropy, Eqn (A2.15) is measuring the likelihood that each state l will transition from an initial state i to a final state f

$$\Delta S = k_B \sum_l \ln\left(\frac{p(r_{f,l})}{p(r_{i,l})}\right) \tag{A2.17}$$

where l now represents an index defining the enclosed sequence length N_l and $l \Leftrightarrow (i,j)$ describes the interaction of monomers i and j (reference 4b). We can integrate the states of l. Exchanging enclosed sequence length ($N(l)$) for the state label l, we obtain

$$\Delta S = k_B \int \ln\left(\frac{p(r_{f,N(l)})}{p(r_{i,N(l)})}\right)dl = \int\left(S(r_{f,N(l)}) - S(r_{i,N(l)})\right)dl \tag{A2.18}$$

Eqn (A2.18) calculates the total change in entropy due to forcing a polymer into a specific configuration that is a function of $N(l)$.

In general, the summation form is much easier to evaluate than the integral form $N(l)$. However, for a RNA chain forming a hairpin in a single stem from 5' to 3' (Figure A2.6), the summation in Eqn (A2.15) can be easily written as an integral. For the GPC with variable γ, Eqn (A2.15) becomes

$$\Delta S = -k_B \sum_{k=0}^{L} \left\{ \gamma \ln[\psi(2k+l)] - \zeta(\gamma,\delta)\left(1 - \frac{1}{\psi(2k+l)}\right)\right\} \tag{A2.19}$$

and, converting to an integral becomes

$$\Delta S \sim -k_B \int_0^L \left\{ \gamma \ln[\psi(2k+l)] - \zeta(\gamma,\delta)\left(1 - \frac{1}{\psi(2k+l)}\right)\right\} dk. \tag{A2.20}$$

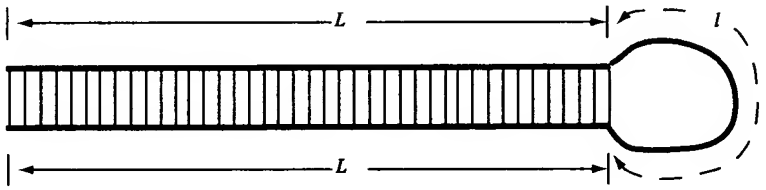

Figure A2.6. Example of a single hairpin containing L base pairs and a loop of length l nt. The total length of the sequence is N.

Now, using the relationship that $N = 2L + l$,

$$\Delta S \sim -\frac{\gamma k_B}{2}\left(N[\ln(\psi N) - 1] - l[\ln(\psi l) - 1]\right) - \zeta(\gamma,\delta)L + \frac{\zeta(\gamma,\delta)}{2\psi}\ln(N/l)$$

$$\sim -\frac{\gamma k_B}{2}\left(N[\ln(\psi N) - 1]\right) \to O\left(\ln(N!)\right) \sim O(N^N). \tag{A2.21}$$

This evaluation can also be done directly. Comparing Eqn (A2.21) with Eqn (A2.3), the result yields an expression of factorial order and $q \propto N$.

We have shown that the entropy must grow factorially with the number of cross-links that are formed. This suggests that there should be little difference between a free chain segment and a stem region except for the fact that the stem has strong correlation while the free strand has comparatively weak correlation mainly introduced by the presence of the stems. This also means that the "penalty" should go mostly to stem formation rather than to loop formation in this model.

Eqn (A2.21) is consistent with the fact that the number of ways that N distinguishable particles can be arranged is $N!$. It is also the normalization weight on the Gamma function probability distribution (ref. 4b). For monomers on a polymer chain, this rule also applies because they (in principle) can be indexed and there is no restriction (in principle) on their arrangements. It is the lattice model that has inadvertently introduced this restriction, particularly when degeneracy is ignored.

This is also consistent with the fact that x-ray diffraction can distinguish these indexed monomers and produces a structural factor proportional to N in the long wavelength regime. Were the true structures that of a lattice, a coordination number (q) should emerge from the lattice parameters and structural factor of the x-ray diffraction data and some of the monomers in the structure would appear to be degenerate. Instead, unique angles are found that are distinguishable (e.g., for proteins, the Ramachandran plots all show non-degenerate distinguishable residues).

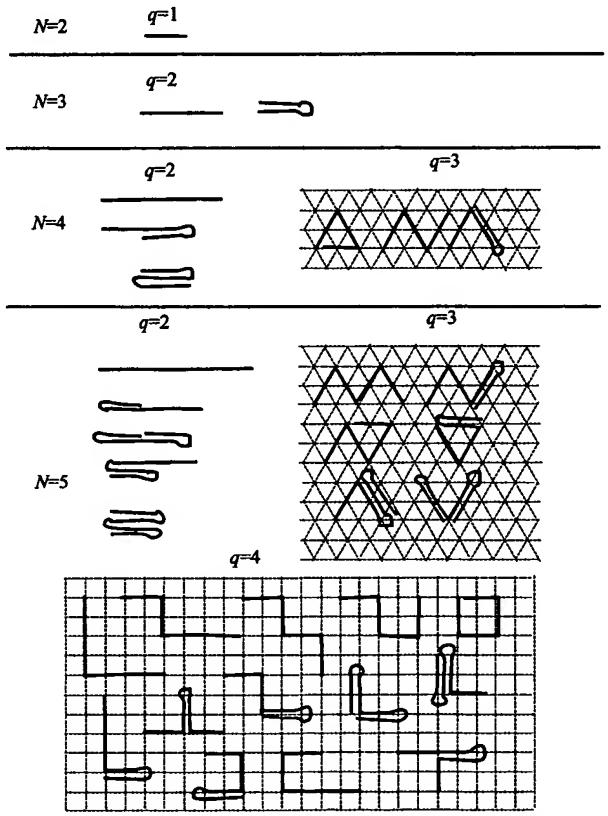

Figure A2.7. An example of one way to draw and classify the smallest number of configurations on a square and triangle lattice. The use of a lattice and, in particular, the arrangement need not conform to this highly restricted form. What is important is that the coordination number is not constant and reaches a similar order as the number of residues in the sequence.

An example of how to classify the conformations of a sequence is shown in Figure A2.7. Here, the growth is $(N-1)!$ and the unique structures are tabulated by the set of coordination numbers $q = 1,2,\ldots,N-1$. This is certainly not the only way to arrange the lattice. The important point is that the coordination number is actually a collection in which the largest one is of the same order as the number of residues.

The model has been derived previously using the assumption of the subadditivity of entropy, and has also been derived by assuming that each connection leads to the creation of a new loop.[4a] A similar approach as outlined here can be used from the stand point of diffusion with the same conclusion. Diffusion is not determined only by the end-to-end separation distance, it is determined by all parts of the polymer chain that diffuse. All that the GPC equation describes is the state r_{ij}. As r_{ij} diffuses, other parts at $r_{i'j'}$ of the structure need not necessarily follow it. Consistent with the contact order, if $i < i' < j' < j$, the rate determining state is r_{ij}.

Finally, this model removes the inconsistencies pointed out at the beginning of this document. A unique value for the coordination number is always obtained with a GPC: $q \sim N$. What is missing in the lattice model approximation is the *degeneracy* of the conformations that are projected onto a lattice structure when one is modeling a freely jointed polymer chain.

References

1. Guth E and Mark H (1934) *Monatsh.* 65: 93. Cited in Ref. 3.
2. Kuhn W (1934) *Kolloid Z.* 68: 2. Cited in Ref. 3.
3. Flory PJ (1953) Principles of Polymer Chemistry. Cornell University Press, Ithaca.
4. a) Dawson W, Suzuki K and Yamamoto K (2001) A physical origin for functional domain structure in nucleic acids as evidenced by cross-linking entropy: parts I and II. J. Theor. Biol. 213: 359-86 and 387-412; b) Dawson, W., Fujiwara, K., Futamura, Y., Yamamoto, K., and Kawai, G. (2006) A method for finding optimal RNA secondary structures using a new entropy model (vsfold). *Nucleosides, Nucleotides, and Nucleic Acids* 25, 171-189.
5. Richardson JS (1981) The anatomy and taxonomy of protein structure. Advances in Protein Chemistry 34: 167-339.
6. Cohen FE, Sternberg MJ, and Taylor WR (1982) Analysis and prediction of the packing of alpha-helices against a beta-sheet in the tertiary structure of globular proteins. J. Mol. Biol. 156: 821-62.
7. Richardson JS (1977) beta-Sheet topology and the relatedness of proteins. Nature 268: 495-500.
8. Adzhubei AA and Sternberg MJ (1994) Conservation of polyproline II helices in homologous proteins: implications for structure prediction by model building. Protein Sci. 3: 2395-2410.

Appendix 3: Deducens machinam in perpetuum moventem ex exemplari[1]

Wayne Dawson

A3.1. Introduction

This document is meant to examine the use of the standard entropy model used in all of the current genre of RNA secondary structure prediction approaches (except Dawson et al.) and in the protein topology problem as proposed by Alm *et al.* The focus is mainly on RNA secondary structure. However we show that as the limits of the model are examined, it increasingly leads to an undesirable prediction for protein and RNA.

A3.2. Distinction between RNA and protein secondary structure

First, we must distinguish between these two different terminologies.

Protein secondary structure refers to a regular arrangement of the polypeptide chain without reference to the side chain (the particular name of the amino acid) or the relative spatial arrangement of these regular arrangements of the polypeptide chains with respect to each other. Therefore, there is no topology information and no indication of how the folded protein is arranged spatially when reference is only made to the secondary structure of a protein.

RNA secondary structure refers to the arrangement of base pairing in the RNA structure. The base pairing defines a relative spatial arrangement between different nucleic acids in the RNA sequence. Therefore, RNA secondary structure does provide essential information on the topology and the spatial arrangement of the RNA sequence.

RNA secondary structure also has a further restriction in that this base pairing arrangement is usually restricted to narrow subset of possible base pairing patterns. The rule is as follows. Let a given RNA sequence be numbered from 1 to N, where N is the total number of nucleotide (nt) in the sequence, and let a given set of distinct base pairs (i,j) and (i',j') be defined such that $i \neq j \neq i' \neq j'$, $i<j$, and $i'<j'$; then the base pairs (i,j) and (i',j') are called RNA secondary structure if they satisfy one of the following conditions: $i<i'<j'<j$, $i',j'<i$, $j<i',j'$ or $i'<i<j<j'$. Hence, typical beta sheet structures that can involve $i<i'<j<j'$ or $i'<i<j'<j$ are not considered in the standard definition of RNA secondary structure. These later structures are called pseudoknots and knots (for RNA structures).

This latter point does not impair the conclusion as applied to proteins. The model can be generalized to account for the entropy in beta sheets in the same way it would be for calculating pseudoknots. Rather, the distinctions should be clearly understood so that the descriptive terms used in this document are not as easily confused.

A3.3. The standard model: the loop penalty model (LP-model)

To this day, in RNA/DNA, the entropy-loss due to folding is evaluated by a topologically *local* function derived from the Jacobson-Stockmayer (JS) equation

[1] Deducing a perpetual motion machine from the model

$$\Delta S(n) = -A - \gamma k_B \ln(n) \qquad \text{(A3.1a)}$$

where A is a fitted constant, k_B is the Boltzmann constant (1.98 cal/mol) and $\gamma(=1.75)$ is the weight that approximates the statistical characteristics of a self-avoiding random-walk where the walker must avoid points that have already been crossed in previous steps. For hairpin loops (H-loops: Fig. 1a, blue region), A is approximately 8.87 cal/mole·K and $n = j - i - 1$ is the enclosed free single strand sequence length (Fig. 1a). For I-loops, $n = n_1 + n_2$ where $n_1 = p - i - 1$ and $n_2 = j - q - 1$ (Figs. 1b and c, blue region) and similarly for bulges. For MBLs, an approximation is used

$$\Delta S = -a - b\sum_i n_i - cs \qquad \text{(A3.1b)}$$

where a, b, and c are all fitted parameters, n_i is the length of the free-strand segments of the MBL (Fig. 1d, blue region), and s is the number of branches. Branches consist of the stems that extend off from the loop.

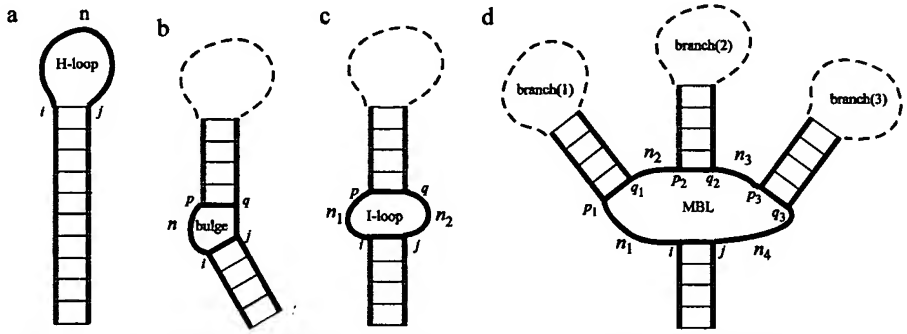

Figure 1: Examples of secondary structure and the corresponding notations. (a) A simple hair pin loop, (b) a bulge, (c) a interior loop, (d) a multibranch loop. The parameter n and ni refer to the length of the given loops (blue).

Models conforming to Eqns (A3.1a) and (A3.1b) assign penalties as a function of the total length of the free-strand segments enclosed by a given loop and by the type of loop that is formed. These structures are topologically local because their entropy only depends on this free strand length in the immediate vicinity (Figs. 1a-d, blue regions) and the entropy can be decided independent of any information about the detailed secondary structure enclosed within (i, j). We therefore call this approach the loop penalty model (LP-model).

It is important recognize that the free energy (FE) is divided into two independent parts in calculating the free energy of RNA secondary structure. The first is the formation of base pairs in a stem and any possible terminal dangling bases that are partially stacked in double strand RNA (dsRNA). The second part is the "penalty"; what

we refer to as the loop-penalty model (LP-model). The penalties are applied in an additive fashion independent of the base pairing free energies to any parts of the structure that are unpaired and closed at both ends to form a loop (the structures show in Fig. 1 where the loop is indicated by the blue region).

In the application by Alm et al., only the H-loop part of the LP-model (Eqn (A3.1a)) is employed (consider Figs. 2ace). The model is not particularly general in this respect.

A3.4. Inspecting the standard (loop penalty) model

For RNA, the loop penalty (LP) model or penalty model consists of base stacking rules that are constant for a given temperature, and a variable function referred to as the Jacobson-Stockmayer (JS) equation: Eqns (A3.1a) and (A3.1b). Alm et al use a similar expression in their Eqn 1.

From thermodynamics we know that all pathways are equally possible. Therefore, though the thermodynamic probability is small that the RNA should fold first from the 5' and 3' ends first (as in Fig. 2b) and work backwards until an entire stem is formed in the shape of Fig. 2a, this pathway is "possible". This property is true whether the sequence length is 20 nt or 20 sextillion nt. Regardless of the sequence length, the polymer "knows" its ends. The same fundamental principles for beta strands of a protein (Fig. 2c and d) must apply to proteins. Indeed, by freely applying the LP-model in their methods, Alm *et al.* admit that these rules must apply.

The loop penalties are a measure of the entropy loss due to folding. Suppose we measure the sequence $A_{100}C_4U_{100}$ at exactly the temperature (T_c) where $\Delta G_{\text{bp}}^{\text{AU}} \equiv 0$ [kcal/mol]. From this simple pattern (neglecting the terminal mismatch free energy at the edge of the loop which are of very similar magnitude), we can write the main contributions to this model as

$$\Delta G(T) = n_{bp}\left(\Delta H_{\text{bp}}^{\text{AU}} - T\Delta S_{\text{bp}}^{\text{AU}}\right) + T\left(A + \gamma k_B \ln(n)\right) \tag{A3.2}$$

where n_{bp} is the number of base pairs (bp), $\Delta H_{\text{bp}}^{\text{AU}}$ is the enthalpy of base-pair stacking, $\Delta S_{\text{bp}}^{\text{AU}}$ is the entropy of stacking, n is the length of the loop, T is the absolute temperature, and A is an empirical constant (Eqn (A3.1a) and (A3.1b)). The first term on the right-hand side we write as

$$\Delta G_{\text{bp}}^{\text{AU}}(T) = \Delta H_{\text{bp}}^{\text{AU}} - T\Delta S_{\text{bp}}^{\text{AU}} . \tag{A3.3}$$

and the second term as

$$\Delta G(L(n),T) = -T\Delta S(L(n),T) = T\left(A + \gamma k_B \ln(n)\right)$$

where $L(n)$ means "hairpin loop of length n".

These simplifications yields

$$\Delta G(T) = n_{bp}\Delta G_{\text{bp}}^{\text{AU}}(T) + \Delta G(L(n),T) . \tag{A3.4}$$

Eqn (A3.3) contains both large entropy and enthalpy terms, but they exactly cancel each other at T_c such that $\Delta G_{bp}^{AU}(T_c) = \Delta H_{bp}^{AU} - T_c \Delta S_{bp}^{AU} \equiv 0$. This permits us to isolate the entropy loss (due to folding) of $A_{100}C_4U_{100}$ from the contributions due to base stacking.

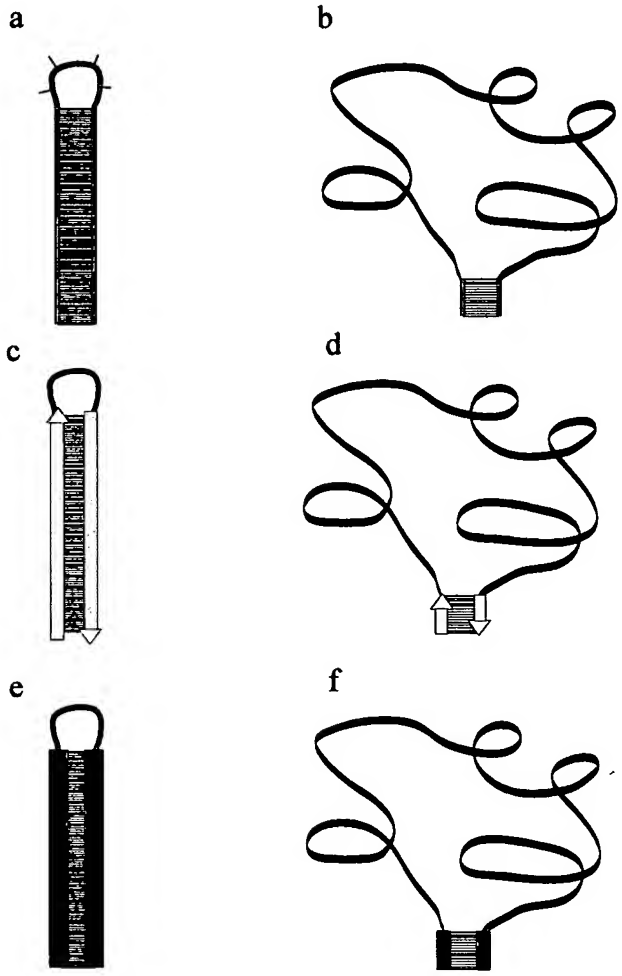

Figure 2. Examples of configurations for an RNA ($A_{100}C_4U_{100}$) or, by analogy, for the topology of two types of protein secondary structures of similar reputed character; one such structure represents a purported ground state, and the other, a short lived structure that exists in the thermodynamic distribution. The RNA sequence would resemble the secondary structures of the following sequences: $G_{100}A_4C_{100}$ and $G_{15}A_{174}C_{15}$ where (a) is the fully folded and (b) is the partially unfolded structure. A combination of two beta strands resembling a fragment from a beta barrel where (c) is the native structure and (d) is the partially unfolded structure and similarly for a combination of alpha helices (e and f).

In the case of protein beta strands, Eqn (A3.3) is not as simple and obvious. First off, we have only limited knowledge on how to make very long beta-strands. Second, it is often believed that the major binding interactions of proteins are due to a hydrophobic

effect; an effect that is largely entropic in nature. Therefore, the temperature behavior will be more complicated. Nevertheless, we can still propose that there exists an amino acid sequence that satisfies the formation of beta-strands that will favorably form a structure such as that shown in Figs. 2c and d. Long beta-sheets are seen in such structures as beta-barrels and these can certainly reach lengths of 20 amino acids. We can therefore propose to borrow a fragment of such a structure as a tangible example for this discussion.

For alpha helices, isolated alpha-helices of such lengths are in principle constructible and, in principle; they could join as shown in Figs. 2e and f. The issue of the attractive interactions will also show similar complexity in the temperature dependence as in the case of the beta-strands.

The base stacking rules of RNA only exist when there is stacking. Granted, after separating, some stacking can occur on a single chain (particularly polyA); however, these corrections are not accounted for in the model. Moreover, corrections for any such details are independent of the stacking free energy ΔG_{bp}^{AU} and, at most, influence the Kuhn length (typically short in single strand RNA: about 5 nt for polyA and 3 nt for poly U). The Kuhn length is not an explicit property used in evaluating the loop penalty model either.[2] All such claims (though bearing consideration) have nothing to do with model that is used, how it calculated, or any corrections that have ever been proposed or made using the model. Nor have these free energy evaluation issues ever been proposed or accounted for as a byproduct of such issues. It follows that they cannot be invoked now and they are not substantially significant even if they are. Therefore, under these specified conditions, we are permitted (by the definitions) to examine the behavior of the loop penalty model in isolation from base stacking rules without claims to anything other than what is *explicitly* in these equations.

We show two configurations for $A_{100}C_4U_{100}$ in Fig. 2. Both Figs. 2a and 2b represent possible but short lived and highly improbable conformations of $A_{100}C_4U_{100}$ at T_c. Their improbability is a function of the entropy loss (due to folding). Since entropy measures the direction the system should go, the expected entropy loss should be such that $-T\Delta S_{\text{Fig(2a)}} > -T\Delta S_{\text{Fig(2b)}}$ because there is more configurational order and restriction in Fig. 2a.

Let $\Delta G_{L(n)} = -T\Delta S_{L(n)}$ be the FE contribution to the loop penalty (LP) for a loop of length n. For simplicity, we assume $T_c = T_{37}$ (37°C) such that the stacking parameters yield $\Delta G_{bp}^{AU} \equiv 0$. Using the true temperature will not change any conclusions. According to the LP-model, at T_{37}, $\Delta G_{L(4)} = 5.6$ kcal/mol and $\Delta G_{L(174)} = 12.2$ kcal/mol (at 37° C). Yet

$$\Delta G_{L(174)}(= -T\Delta S_{\text{Fig2b}}) > \Delta G_{L(4)}(= -T\Delta S_{\text{Fig2a}}). \qquad \text{(A3.5)}$$

---

[2] I have shown in Dawson et al 2001 that the cross linking entropy model can match the Jacobson-Stockmayer expression under limiting conditions. The JS expression was originally fitted to these conditions with experimental data. The CLE-model considers the Kuhn length whereas it is a hidden parameter in the JS expression. Given sufficient data, the CLE-model can account for all these issues, such is unlikely for the JS expression.

This is a contradiction.

Neither do we avoid this contradiction for the protein example in Figs. 2c and d, nor that of Figs. 2e and f. The same rules must apply for proteins. In Eqn 1 of Alm et al, the entropy-loss is figured in the same way, and we can propose some temperature T_c where the attractive interactions between the beta-strand (Fig. 2cd) or alpha-helices (Fig. 2ef) become too weak to hold the structure together. The first term in Eqn 1 of Alm *et al.* is worked out as an additive function and there necessarily exists some temperature where this expression goes to zero.

Returning back to examining the issue in terms of RNA, more important is the direction. The difference in the free energy is

$$\Delta G_{L(4)} - \Delta G_{L(174)} = 5.6 - 12.2 = -6.6 \text{ kcal/mol} \qquad \text{(A3.6)}$$

and because this is negative, the direction is toward Fig. 2a and away from Fig. 2b even though the latter should have less structural order. If we work basepair-by-basepair to $L(4)$ or in chunks $L(174) \rightarrow L(164) \ldots \rightarrow L(4)$, the difference in the free energy (as in Eqn (A3.6)) is always negative. In other words, although infrequent, the LP-model predicts that we should encounter Fig. 2a for $A_{100}C_4U_{100}$ far more frequently than Fig. 2b and the system will move away from Fig. 2b *toward* Fig. 2a. Strangely, only one bp formed at $L(4)$ costs the same price as forming all of the base pairs for the structure.

Since the phenomenon does not employ any stacking whatsoever, this property should also work for non-Watson-Crick sequences such as polyA. This would permit us to remove all temperature considerations from the calculations and run this experiment independent of temperature from above freezing to where the sample degrades.

This contradiction (*were it true*) permits us to violate the second law of thermodynamics. First, we biotinylate the $5'$ and $3'$-ends of the sequence $A_{100}C_4U_{100}$ and force the $5'$ and $3'$-ends toward each other. The structure will spontaneously collapse into the structure similar to Fig. 2a if we do work greater than $\Delta G_{L(4)} = 5.6$ kcal/mol on the system, where $L(4)$ is the minimum allowed loop size and $\Delta G_{L(4)}$ is the approximate FE penalty. It is spontaneous because the difference in the FE is negative: $\Delta G_{L(4)} - \Delta G_{L(174)} = 5.6 - 12.2 = -6.6$ kcal/mol. The structure will collapse because any work we apply that exceeds $\Delta G_{L(4)}$ will drive the system toward the minimum configurational entropy loss and this happens to be a loop structure similar to Fig. 2a (according to the LP-rules). So first, it has generated heat rather than required work yet produced a structure with far greater order. Now we simply release our applied force, and the structure goes through isothermal expansion to its equilibrium structure doing $\Delta G_{L(174)} \approx 12.2$ kcal/mol of work. This is because (1) we biotinylated the $5'$ and $3'$ ends so the structure will do work against that and (2) the *difference* between the entropy-loss of the structure in Fig. 2b ($-A - k_B\gamma \ln(174)$) and the denatured structure (0, set by definition) is then $\Delta G_{L(184)}$. These were the definitions.

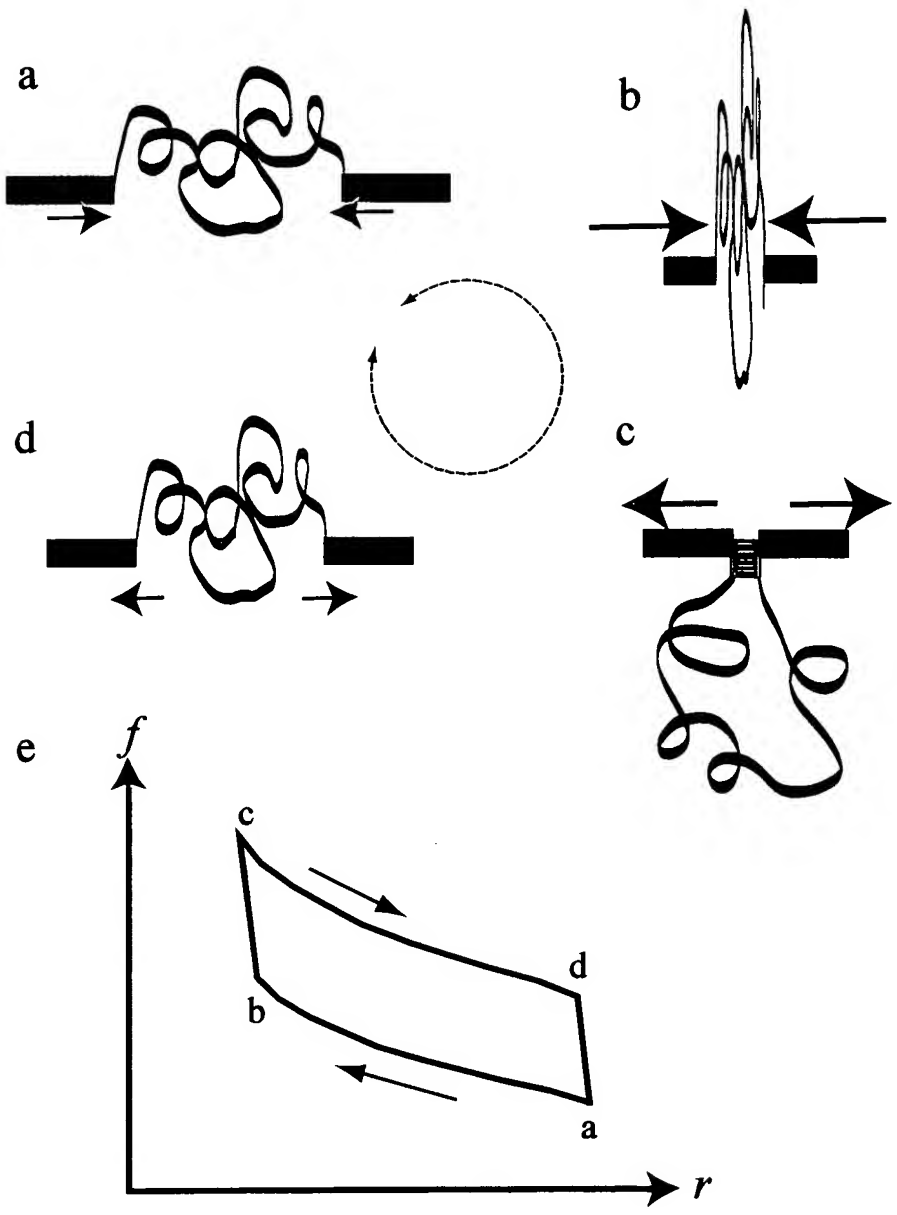

Figure 3. The correct thermodynamic cycle for a polymer model. (a) Work is done to press the 5'-3' ends together (orange arrow). (b) Here, the structure makes an adiabatic transition ($T_a \rightarrow T_c$) to (c). (c) Releasing the work done on the ends, the structure now behaves by pushing out on the levers. (d) An adiabatic transition to (a); the polymer has pushed out until it returns to the original equilibrium distance in (a) warming up in the process $T_c \rightarrow T_a$. (e) An fr (pV) diagram of what is happening.

We put in 5.6 kcal/mol of energy (heat) and got back 12.2 kcal/mol of work. We gained 6.6 kcal/mol of excess work that we can extract to do useful work. This is not a big gain, but it could be improved upon by turning to non-Watson-Crick type structures

such as polyA, making the sequence even longer (*e.g.*, A_{10000}), adding more levers to form internal loops and coupling many of these structures together. Indeed, we could use DNA and add even further protection against degradation, thereby extending the material's lifetime. It's well worth the development because it would be possible to heat an entire skyscraper by coupling many of these together and placing them in separate rooms. In fact, we could use this process to feed storage batteries that could be used to power the transportation system. That would surely solve the problem of global warming due to greenhouse gases and many of the world's energy problems too.

A3.5. Comparing the LP- and CLE models: the Carnot engine

We now compare the Carnot engine for an ideal gas with the model above.

From the ideal gas equation we have

$$TdS = dU + pV \tag{A3.7}$$

where dU is the change in internal energy and $W = pV$ is the work done by the as. The work done during isothermal expansion or compression ($dT = 0$) of an ideal gas is

$$\Delta W = -\int pdV = -\int_{V_1}^{V_2} \frac{nRT}{V} dV = -nRT \ln(V_1 / V_2). \tag{A3.8}$$

During the adiabatic expansion or compression cycle of the ideal gas ($TdS = 0$) we have

$$TdS = 0 = dU - dW = c_v dT - dW \quad \text{or} \quad \Delta W = c_v \Delta T \tag{A3.9}$$

where c_v is the specific heat at constant volume. From the above expressions

$$c_v dT + \frac{nRTdV}{V} = 0 \quad \text{leading to} \quad \int c_v \frac{dT}{T} + \int \frac{nRdV}{V} = 0 \tag{A3.10}$$

which, for adiabatic expansion or compression of a gas, yields

$$T_2 = T_1 \left(\frac{V_1}{V_2} \right)^{\gamma - 1} \tag{A3.11}$$

where $\gamma = c_p / c_v$.

With the polymer equations, it is known that c_r (analogous to c_v) is not so large because it is generally viewed that $\Delta U \square 0$ in the ideal polymer. However, it is not zero, and so we write the adiabatic case in a similar fashion

$$c_r dT + f dr = 0, \text{ which leads to } c_r \frac{dT}{T} - \left(\frac{\partial S}{\partial r} \right) dr = 0 \tag{A3.12}$$

and this in turn leads to

$$c_r \ln\frac{T_2}{T_1} + \left(\Delta S(r_2) - \Delta S(r_1)\right) = 0 \qquad \text{(A3.13)}$$

where T_1, r_1 refer to one state on the fr-graph in T_2, r_2 refer to a second state. These states are shown in Fig. 3e and the corresponding RNA structures associated with them are shown in Fig. 3a-d.

A3.5.1 The fr-diagram for the CLE model

Inserting the CLE expression in to Eqn (A3.13)

$$c_r \ln\frac{T_2}{T_1} \approx -k_B\left\{2\gamma \ln\frac{r_1}{r_2} - \zeta\left(1 - \frac{r_2^2}{r_1^2}\right)\right\} \qquad \text{(A3.14)}$$

which indicates that the polymer cools off when $r_1 > r_2$ and therefore $T_2 < T_1$.

Since, c_r is rather small, this means that there is very little adiabatic interaction for the polymer as it moves from the two temperature baths shown in Fig. 3. However, there would be a change in temperature when the structure becomes more ordered: the structure would cool down. Since r_a corresponds to state 1 and $r_a > r_c$, the temperature at r_c will have decreased according to Eqn (A3.14). Likewise, when the force is released at C, the structure will begin to expand, and, since r_c corresponds to state 1, and $r_c > r_a$, the system will warm up again. If this is a true Carnot engine, then the process is completely reversible and there is no gain in useful work: the system returns to its original configuration and the work put in is close to the work returned (assuming all the frictionless, lossless, etc. conditions apply to the source of work put into the system, which is clearly unrealistic). The CLE model is entirely consistent with the basic thermodynamics that are expected of an idealized system.

A3.5.2 The fr-diagram for the loop penalty model

For the loop penalty model, the process is shown in Fig. 4. If the JS model is the case in the expected scenario of a folding polymer, we would push the ends together till we reached something that looks a little like an unbound state C, then spontaneous collapse would drive the structure from Fig. 4X to C in an irreversible process collapsing the structure into the most ordered configuration. In Fig. 4e, this is represented by the dotted blue line jumping from position X to C and a break in the fr diagram in Fig. 4e. The system would cool slightly during the transition from X to C according to Eqn (A3.14) because $r_a > r_x$ and therefore we would have a drop in temperature to T_x from T_a. However, note that $T_x > T_c$. Therefore, the cooling down does not reach the same magnitude as T_c, and this means we have extra heat that should not be there. The force is now released and the polymer now expands as it passes to from C to D in Fig. 4e. An adiabatic effect occur in the cycle at the transition from

D to A with the system warming up because here $r_c < r_a$ and therefore $T_a > T_c$. However, note that because $T_x > T_c$, we have a net gain in temperature (T_g) equal to $T_g = T_a - T_x$. The work done by the polymer is via $dw = fdr$. Because of the jump at X, the work done *on* the system is less than the work done *by* the system.

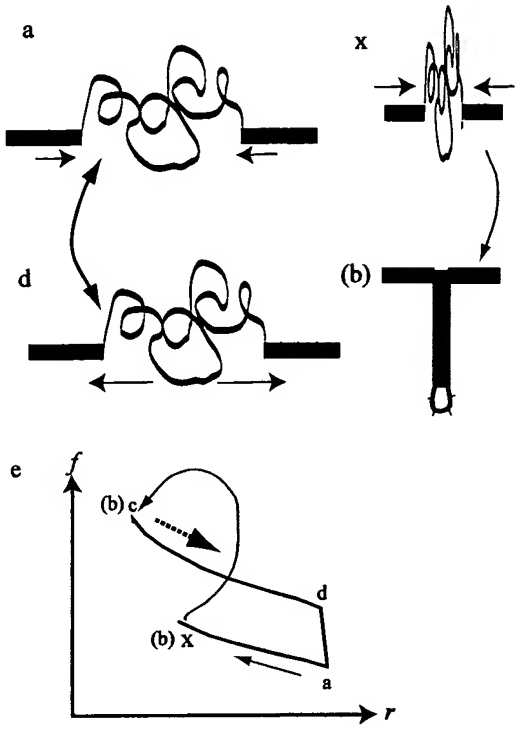

Figure 4. The incorrect thermodynamic cycle generated by the LP model. (a) Work is done to press the 5'-3' ends together (orange arrow). (x) Before finishing pressing the 5'-3' ends completely together (as in Figs. 3b and c), the structure spontaneously collapses into the structure at c with $T_a \rightarrow T_x$ where T_x is the effective temperature associated with position r_x. The dotted arrow indicates that system makes a non-reversible jump to c. (c) Releasing the work done on the ends, the structure now behaves by pushing how on the levers. (d) Because we have the 5'-3' ends connected, the structure responds as though it were structure (c) and pushes out until it returns to the original equilibrium distance in (a). (e) An fr (pV) diagram of what is happening.